Write-Up
Project 2-Group 16

Isbelis Castro, Ben Rosensweig, & Lindsey Krempa
July 23th, 2024

Introduction

Project 2 required our group to apply what we have learned about Python, Pandas, and PostgreSQL to build an ETL pipeline that extracted and transformed data from multiple CSV files. This collaborative process proved to have valuable real world applications because we cleaned and reconstructed data that can be used for purposeful analysis. Specifically for Project 2, we extracted and transformed crowdfunding Excel data that can be used for meaningful analysis.

Our group was able to generate an ERD after extracting and transforming the data. The ERD organized the campaign, contacts, category and subcategory dataframes. This ERD was then used to create table schemas for the CSV files. Creating the schemas required us to consider and specify data types, primary keys, foreign keys, and other constraints. These schemas were then used to produce tables in Postgres, allowing us to load the extracted and transformed data.

Extract/Transform/Load code overview (not line by line, just broad ideas)

In this project, working on a crowdfunding database involves extracting, transforming, and loading data to create a comprehensive database for analysis. First, using Python and Pandas, data was extracted from various sources such as CSV files and databases and loaded into Pandas DataFrames. This involved reading CSV files and connecting to databases. During the transformation phase, data was cleaned to handle missing values, remove duplicates, and ensure consistency. Necessary transformations, such as formatting dates and creating new features, were applied, followed by merging data from different sources to enrich the dataset.

The data cleaning process included several steps. For example, the **category** and **subcategory** columns were split, and the original column was dropped. Unnecessary columns were removed, and the **blurb**, **launched_at**, and **deadline** columns were renamed. The **goal** and **pledged** columns were converted to float data types, and datetime columns were appropriately formatted. In the contacts DataFrame, columns were transformed using regular expressions to manipulate string data. The transformed DataFrame was saved with dropped and reordered columns.

After transformation, the data was loaded into a SQL database, creating tables for categories, subcategories, campaigns, and contacts. The database schema was defined, and an Entity-Relationship Diagram (ERD) was created to visualize the relationships between these

tables. The ERD was generated using QuickDatabaseDiagrams (QuickDBD), and the schema was imported into a PostgreSQL database.

For analysis, queries were performed to determine the frequency of campaign outcomes (successful, failed, live, etc.) and visualized using bar charts. Analysis of categories and subcategories was conducted using bar charts to identify the most successful campaigns. The average number of backers for successful vs. failed campaigns was analyzed using regression and scatter plots. Python libraries like Pandas and NumPy were used for data manipulation and analysis, while Matplotlib and Seaborn were used for visualizations. For example, to understand the number of successful versus failed campaigns, a bar chart was created. A donut chart and a bar chart visualized which categories and subcategories had the most successful campaigns, respectively. To compare the average number of backers for successful versus failed campaigns, a scatterplot was used. These analyses and visualizations provided insights into the performance and characteristics of crowdfunding campaigns.

Analysis

For our second query, we wanted to compare the number of successful and failed campaigns in our database. Overall, the successful campaigns were the most frequent with 565, following failed campaigns (364), canceled campaigns (57), and live campaigns. We created both a bar chart and a donut chart to compare the frequency of each campaign outcome. We initially decided on creating the bar chart, but then ultimately decided on the donut chart as well since the bar graph was needed to properly visualize our second query that focused on categories and subcategories.

Queries 3-7 analyzed the category and subcategories with the most successful campaigns and other related variables. Follow the analysis below:

Geographical Distribution: The US is the predominant location for crowdfunding campaigns, representing 80.6% of the total. This high frequency underscores the significant role of the US in the crowdfunding landscape. (see graph: Number of Campaigns by Country).

The Most Successful Category and Subcategory Theater stands out as the most successful campaign category in the crowdfunding database, accounting for 33.1% of all campaigns and including the subcategory Plays. In contrast, Journalism campaigns have the lowest success rate during the analysis period. The second most successful category is Film & Video, with an 18.1% success rate, particularly strong in the Documentary subcategory, which has a 33.7% success rate (n=60). Music ranks third, with a 17.5% success rate, primarily driven by the Rock subcategory, which boasts a 48.57% success rate (n=85) (see graphs: Most Successful Campaigns by Category and Most Successful Campaigns by Subcategory).

Successful vs. Failed Campaigns The bar chart "*Monthly Outcome of Campaigns*" shows that successful campaigns occur more frequently than failed ones, with the peak occurring in January 2021. Interestingly, all campaigns in January 2020 failed, while all campaigns in

February 2022 succeeded. The graph "Yearly Outcome of Campaigns" reveals that 2022 had the fewest campaigns, while 2021 saw the highest number of both successful and failed campaigns.

Success Rate by Category The box plot "*Success Rate by Category*" reveals that Technology has the highest success rate, exceeding 20%, followed closely by Theater at around 18%. Additionally, the categories of Food, Technology, Theater, and Games show a positive skew, with Food exhibiting the most variability.

Goal by Category The box plot "*Goal by Category*" highlights that the categories with the greatest variability are Theater, Film & Video, Publishing, and Games. All these categories are positively skewed, except for Photography and Journalism.

Pledged by Category The box plot "*Pledged by Category*" indicates that Theater is one of the most variable categories, with no outliers. Similar to the analysis of goals by category, all categories are positively skewed, except for Photography and Journalism.

Correlation between goal and number of backers For the scatter plot titled *"Pledge Goal vs Backers for Successful and Failed Campaigns"* Successful campaigns had a higher number of backers by an increased r-squared of .07, from 0.46 to 0.53. In this dataset, success seemed to be guaranteed for campaigns that had both a goal of $50,000 or less, and had at least 500 backers. Any campaign with a goal above $50k and had fewer than 1000 backers seemed doomed to fail. I also wanted to see whether there was a significant difference between the average donation of backers between successful and failed campaigns - I didn't find any. With the average donation of successful campaigns at $69, and at $64 for failed campaigns, campaigns don't succeed or fail depending on the donation sizes of their backers.

Bias/Limitations

One limitation is a lack of data on the rate of fulfillment of a campaign over its duration, or rather, timestamps of each donation. If we knew how many backers were present over the course of a campaign's length of time, we would be able to further predict whether campaigns would succeed or fail depending on the number of backers a campaign had at its outset. For example, if we had this kind of data, we would be able to see how quickly successful campaigns fulfilled 25%, 50%, and 75% of their stated goals. A hypothesis that could be explored is whether campaigns that have more backers at the beginning of a campaign, say its first week or month, are more successful than failed ones.

Conclusions/Reflection

This was a project that touched on several aspects of Data Analysis, including ERDs, SQL, SQLAlchemy, and Extraction, Transforming, and Loading data. As a final summation of findings from this dataset, the US had the highest frequency of campaigns, with successful campaigns

occurring more frequently than failed ones, with notable peaks in January 2021 and with the most failures in January 2020. Goals and pledged amounts exhibit variability across most categories. The category of Technology leads with the highest success rate. This analysis can guide future campaign strategies and offer insights into the most promising areas for crowdfunding initiatives.